Are Points in Tennis Independent and Identically Distributed? Evidence From a Dynamic Binary Panel Data Model

Franc J. G. M. KLAASSEN and Jan R. MAGNUS

This article tests whether points in tennis are independent and identically distributed (iid). We model the probability of winning a point on service and show that points are neither independent nor identically distributed: winning the previous point has a positive effect on winning the current point, and at "important" points it is more difficult for the server to win the point than at less important points. Furthermore, the weaker a player, the stronger are these effects. Deviations from iid are small, however, and hence the iid hypothesis will still provide a good approximation in many cases. The results are based on a large panel of matches played at Wimbledon 1992–1995, in total almost 90,000 points. Our panel data model takes into account the binary character of the dependent variable, uses random effects to capture the unobserved part of a player's quality, and includes dynamic explanatory variables.

KEY WORDS: Binary choice; Dependence; Dynamic panel data; Linear probability model; Nonidentical distribution; Random effects; Tennis.

1. INTRODUCTION

This article attempts to model the probability of winning a point on service in professional tennis. This probability is the key to most analyses in tennis. For instance, it provides the basis for the calculation of the probability of winning a match. In most work on tennis, points are assumed to be independent and identically distributed (iid), which implies that the key probability is constant for a player throughout a match.

In this article we test the iid hypothesis and reject it. We present a model that explicitly captures both the dependence (measured by the impact of the previous point) and the nonidentical distribution (measured by the "importance" of the current point) and use it to analyze the deviations from iid in more detail. We find that weaker players violate the iid hypothesis more than stronger players. Deviations from the iid hypothesis, although statistically strongly significant, are not large, and thus the hypothesis will serve as a reasonable first-order approximation in many applications.

As a preview of our results (fully reported in Sec. 4.3), consider a match in the men's singles (women's singles) between two average players. Overall, these players will win 65% (56%) of their service points. If the previous point was won (and if the current point is not the first point in the game), then the probability of winning a point increases by .3% (.5%), reflecting a "winning mood." However, if the previous point was lost, then the probability of winning a point decreases by .5% (.7%). Also, the more "important" (defined in Sec. 4.2) the point, the lower the probability that the server wins the point. For example, at a point of zero importance, the probability of winning a point on service increases by .4% (.6%) compared to a point of average importance. At 30–40 (break point) in the first game of the match, the probability of winning a point on service decreases by .8% (.8%); at 5–5 in games and 30–40 in points in the first set, it decreases by 1.5% (2.1%); and at 5–5 in games and 30–40 in points in the fifth (third) set, it decreases by 4.6% (4.8%).

To test the iid hypothesis at point level and to analyze differences across players, we need point-to-point data over many matches involving different players. We were fortunate in obtaining point-to-point data on 4 years of Wimbledon men's and women's singles, 1992–1995, distributed over 481 matches, leading to 57,319 points in the men's singles and 28,979 points in the women's singles. This is one respect in which our article differs from the existing literature on the statistical analysis of tennis, which is hampered by a serious lack of detailed data. If some data are available, they are either based on end-of-match results (6–4, 6–3, 6–3 say) of several matches (Croucher 1981; Jackson and Mosurski 1997), or occasionally on a point-to-point analysis of one match, usually an important final (Croucher 1995). This article is the first one using data for many matches and at point level. The data concern only Wimbledon, one of the tournaments played on fast grass courts, and the generality of our conclusions may be restricted by this fact.

One characteristic of our dataset is that it involves heterogeneous players. If one estimates the probability of winning a point on service using pooled data without a proper correction for the quality of players, then one will find that winning the previous point has a positive impact, even if points for each player individually are independent. The reason for this effect is that winning the previous point contains a small but positive piece of information about a player's quality. We call this "pseudodependence," and it should be carefully distinguished from true dependence (the kind we are interested in): the presence of an effect of the past on the current point for one player.

In order to distinguish "pseudodependence" from true dependence, we correct for the quality of a player. But only part of that quality is observable (e.g., the player's ranking),

Franc Klaassen is Assistant Professor, Faculty of Economics and Econometrics, University of Amsterdam, Roetersstraat 11, 1018 WB Amersterdam, The Netherlands (E-mail: *klaassen@fee.uva.nl*). Jan Magnus is Professor of Econometrics, CentER, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands (E-mail: *magnus@kub.nl*). The authors are grateful to IBM U.K. and The All England Lawn Tennis and Croquet Club at Wimbledon for their kindness in providing the data, to Katia Zhuravskaya for excellent programming assistance at the early stages of this project, and to Maurice Bun, Erwin Charlier, Bas Donkers, Frank Kleibergen, and three referees for very useful and constructive comments.

whereas the rest is unobservable ("form of the day," fear against a specific opponent). This is our first problem. To solve this problem, we model the unobserved part of quality as a random individual effect, in the same spirit as one typically corrects for unobserved heterogeneity in panel data (Hsiao 1986, p. 33). Hence we explicitly use panel data techniques to correct for pseudodependence and view our dataset as a panel—that is, a cross-section of matches where each match comprises two time series of service points, one for each player.

Although panel data techniques are appropriate for our tennis data, direct application involves two additional problems. First, our focus variable (winning a point on service) is a binary variable—it can only take the values 0 (if the point is lost) and 1 (if the point is won)—whereas standard panel techniques are designed for continuous data. Second, most of our regressors are dynamic; for example, the performance on previous points, which captures dependence, is a function of lagged values of the dependent variable. The usual panel estimators, such as the within-group estimator, then become inconsistent (Hsiao 1986, p. 72).

The estimation of dynamic panel models with a discrete dependent variable is essentially an unsolved problem in the classical statistics literature, although partial solutions exist; for example, the logit approach by Honoré and Kyriazidou (2000). The Bayesian literature (Albert and Chib 1993; Chib and Greenberg 1994; Johnson and Albert 1999) also provides solutions for the probit and other link functions. Moreover, the Bayesian approach is exact rather than asymptotic. However, our very large dataset makes the latter advantage irrelevant and, more important, makes the computational burden in the Bayesian approach excessive.

To solve the estimation problem, we exploit the special nature of our tennis data. This allows us to use the linear probability model to obtain results very similar to probit and logit (Sec. 3). Thus we can use standard dynamic panel techniques (Baltagi 1995, chap. 8; Hsiao 1986, chap. 4; Mátyás and Sevestre 1992, chap. 6). We estimate our model by feasible generalized least squares (FGLS), which we show to be consistent for our dynamic panel, taking into full account the effects of the binary structure on the first two moments of the observations and the nonobservables. Hence we provide a practical solution to the estimation of binary dynamic panels, and illustrate its effectiveness by applying it to our Wimbledon data.

In the statistical literature on tennis, only a small number of works are concerned with some aspect of independence and identical distribution of points. Croucher (1981), studying the "back-to-the-wall effect" (the possible effect that a player or a team plays better if trailing) in tennis, found only very slight evidence of this effect. Jackson and Mosurski (1997) investigated whether "getting slammed during your first set might affect your next." In other words, they challenged the independence assumption and concluded that there is dependence, possibly caused by "psychological momentum." In earlier work (Magnus and Klaassen 1999a–c) we tested 11 tennis hypotheses, many of them relating to the iid hypothesis, and rejected most of them.

The question of independence and identical distribution has been a hot topic in other sports, particularly basketball and baseball. Dependence between points is called "streaks" in baseball. Lindsey (1961) found that the distributions of runs scored in different half-innings in baseball are not homogeneous (the first and third half-innings have the highest expected number of runs). Nevertheless, Lindsey concluded that scores can be replicated assuming independence of runs scored in different innings. Simon (1971, 1977) noticed that of the 31 World Series played from 1945–1975, 18 have lasted seven games (the maximum). From this he concluded that there must be a back-to-the-wall effect in which the team that is behind performs better, thus challenging the independence assumption. Siwoff, Hirdt, and Hirdt (1987, p. 97) found that the probability of hitting well in a game is independent of whether or not the hitter is on a streak. Albright (1993) also did not find convincing evidence of streaks. However, commenting on Albright's article, Stern and Morris (1993) and Albert (1993) suggested alternative approaches that might lead to a different conclusion. Stern (1995) suggested, however, that streaks might exist. Thus the question remains unresolved.

In basketball, dependence between points is known as the "hot hand." Research on the hot hand started with psychologists Gilovich, Vallone, and Tversky (1985), who concluded that people believe in the hot hand (not that a hot hand actually exists). Larkey, Smith, and Kadane (1989) examined game data for 18 NBA players, in particular Detroit's Vinnie Johnson, who has a reputation of being one of the most streaky of shooters. From their descriptive analysis they concluded that Johnson is indeed a streaky shooter and hence that the hot hand exists. In contrast, Camerer (1989) showed that although the basketball market believes in the hot hand, in fact there is no such thing in basketball shooting. He found that bets placed on teams with a current winning streak are more likely to be *losers* than winners, and he concluded that teams on winning streaks are mistakenly believed to be hot. In a comment, Brown and Sauer (1993) disagreed with Camerer and showed that hot-hand effects are present. Clearly, this issue is also unresolved.

This article is organized as follows. In Section 2 we briefly discuss the data and the representativeness of the sample. In Section 3 we discuss the various problems that arise in modeling a binary dynamic panel dataset with random effects, and show how this model can be consistently estimated (with details given in the Appendix). Section 3 (and the Appendix) contain the theoretical part of the article and can be read without any knowledge of or interest in tennis. In Section 4 we discuss the choice of variables for our tennis application, present our estimation results, and test the iid hypothesis. We provide extensive sensitivity analyses and diagnostics in Section 5, and a conclusion in Section 6.

2. THE DATA

Our data consist of 481 matches played at the Wimbledon championships during 1992–1995: 258 matches in the men's singles and 223 matches in the women's singles. In each match we know the two players and the complete sequence of points. Because of the special nature of points played in tie-breaks, these points have been excluded from the analysis in this article. Because men play for three won sets and women for two, we have about twice as many points for the men (57,319) than

for the women (28,979). We have described the data in detail in earlier work (Magnus and Klaassen 1999a).

We do not have data on all matches played during the 4 years. In fact, we have only the matches played on one of the five "show courts" (Centre Court and Courts 1, 2, 13, and 14), because these are the only courts where the data collection occurred. As a result, we have data on almost half of the 1,016 matches played. Typically, matches involving the most important players are scheduled on the show courts, and this causes an underrepresentation in the dataset of matches involving weaker players. All results have been corrected for this selection problem by weighting the matches by the inverses of the sampling percentages. We have discussed the (admittedly imperfect) weighting procedure in Magnus and Klaassen (1999b). In fact, whether or not we weight, we still reject the iid hypothesis (see Sec. 5, question 9).

The focus variable in this article is the probability of winning a point on service. Averaging over all players, this probability is estimated as .645 (with a standard error of .002) in the men's singles and .559 (.003) in the women's singles, a large difference. This is one important difference between men's singles and women's singles, but we have also reported other differences (Magnus and Klaassen 1999a–c). As a result, we keep the analyses for the men's singles and the women's singles separate.

3. DYNAMIC BINARY PANEL DATA WITH RANDOM EFFECTS

In this section we develop the model and describe a consistent method of estimation, details of which we provide in the Appendix. The tennis aspect is minimal in this section, so the results should be of more general interest.

We regard our data as a panel consisting of N matches (258 in the men's singles, 223 in the women's singles). We assume throughout that matches are independent, and thus we begin by considering one match.

We wish to test the iid assumption. As a vehicle for doing so, we model the probability of winning a point on service. Thus let y_{at} be 1 if player $\mathcal{A}$ wins his or her tth service point (against player $\mathcal{B}$) and 0 otherwise. (We write y_{at} instead of y_{abt} for simplicity of notation, although y_{at} depends on both $\mathcal{A}$ and $\mathcal{B}$.) Similarly, let y_{bt} be 1 if $\mathcal{B}$ wins his or her tth service point (against $\mathcal{A}$) and 0 otherwise.

Within each match of T points, we have data on T_a service points of player $\mathcal{A}$ and T_b service points of player $\mathcal{B}$. An average match in the men's singles comprises $T = T_a + T_b =$ 222 points (130 points in the women's singles).

The two players $\mathcal{A}$ and $\mathcal{B}$ in each match are modeled symmetrically. Concentrating on player $\mathcal{A}$, our starting point is the linear probability model

$$y_{at} = Q_a + D_{at} + \epsilon_{at}, \tag{1}$$

which comprises three components: quality Q_a, other (dynamic) regressors D_{at}, and random errors ϵ_{at}. (Again, we write, for instance, Q_a instead of Q_{ab} for simplicity of notation.) Equation (1) says that the probability that $\mathcal{A}$ wins the tth service point is equal to the expectation of $Q_a + D_{at}$ (assuming that ϵ_{at} has expectation 0).

The choice of a linear probability model requires a justification. First, because $E(y_{at}) = \Pr(y_{at} = 1)$ because of the binary character of y_{at}, the estimated expectation must lie between 0 and 1. This is not necessarily the case in general, but in our tennis dataset it is always the case. In fact, the estimated probability of winning a point on service for each player lies in the interval (.55, .75) in the men's singles and (.42, .73) in the women's singles; see also Section 5, question 1. Second, the binary character also implies second-moment restrictions on the model. The equality restriction $E(y_{at}^2) = E(y_{at})$ will be imposed, whereas the second-moment inequality restrictions $0 < E(y_{at}y_{as}) < 1$ and $0 < E(y_{at}y_{bs}) < 1$ are satisfied (Sec. 5, question 1). We do not consider higher-order moment restrictions, because our FGLS method of estimation uses only first and second moments. Third, alternative nonlinear models (such as probit and logit) will yield comparable results, because the relative deviation in the range (.35, .70) between the link functions of the linear probability model and the probit and logit models (appropriately centered and scaled) is less than 1% (see Hsiao 1986, p. 155; Maddala 1983, p. 23). Given that the linear and nonlinear models will give similar results, we prefer the linear probability model since it is easier to use than panel logit or probit models and allows us to use standard (and computationally fast) GLS techniques in solving the dynamics problem (Sec. 3.4).

Our approach to the problem is classical, not Bayesian, although Albert and Chib (1993) and Johnson and Albert (1999, chap. 3) showed how to fit probit and other models using data augmentation and Gibbs sampling. In addition, Chib and Greenberg (1994) described Bayesian fitting algorithms for ARMA models that can be extended by data augmentation to discrete response models. We prefer the classical approach for three reasons: our dataset is large, and hence the difference in results between the classical and Bayesian analysis will be small; the computational burden in using the Bayesian approach will be excessive in our case; and, because the linear probability model yields very similar results to probit and logit for the tennis data (see earlier), there is no reason to use the more-demanding Bayesian methods needed to estimate logit or probit models.

We discuss each of the three components Q_a, D_{at}, and ϵ_{at} in (1) in turn. The quality of player $\mathcal{A}$ against player $\mathcal{B}$, denoted by Q_a, is a crucial ingredient, because it enables us to distinguish pseudodependence from true dependence, as discussed in Section 1. Pseudodependence is closely related to the problem of spurious state dependence or heterogeneity bias in the panel data literature. The problem occurs when we ignore heterogeneous intercepts in a pooled regression, and it leads to biased slope estimates (see Hsiao 1986, p. 6). A commonly used method to avoid heterogeneity bias in panel models is to use individual-specific intercepts. In our context of modeling the probability of winning a service point in tennis, Q_a is such an individual intercept. Section 3.1 discusses quality in more detail.

Whereas Q_a contains characteristics of $\mathcal{A}$ and $\mathcal{B}$ at the beginning of the match, the dynamic term D_{at} depends on all match information (of both players) up to but excluding point t. If points were iid, then such match information would be useless in predicting y_{at}. Hence the variable D_{at}

captures departures from the iid hypothesis, such as dependence variables capturing a winning mood (if such a mood exists), and characteristics of the point being played (such as the importance of a point). Section 3.2 provides further details.

The random error ϵ_{at} in (1) is affected by the binary structure of y_{at}, because it can take only the values $0 - Q_a - D_{at}$ and $1 - Q_a - D_{at}$. The implications of the binary structure on the error term are discussed in Section 3.3.

3.1 Quality

The proposed quality variable Q_a contains some components that we observe (most notably the ranking of the two players) and many that we do not observe (such as form of the day, fear against a specific opponent, and special ability, if any, on grass). We assume that observed quality is linear and denote it by $x_a'\beta$. (The specific ranking-based definition of x_a for our tennis data is discussed in Sec. 4.1.) Unobserved quality is denoted by η_a, and we model it as a random individual effect, just as one typically corrects for unobserved heterogeneity in the panel data literature.

Thus motivated, we write quality as

$$Q_a = x_a'\beta + \eta_a. \tag{2}$$

We assume that the observed part contains a constant term, so that there is no loss in generality in assuming that $E(\eta_a) = E(\eta_b) = 0$. In addition, we impose

$$\operatorname{var}(\eta_a) = \operatorname{var}(\eta_b) = \tau^2, \qquad \operatorname{cov}(\eta_a, \eta_b) = \gamma, \tag{3}$$

where $|\gamma| < \tau^2$. The assumption that the variance τ^2 of the random effect is constant across players is standard (Hsiao 1986, p. 33), but the introduction of a covariance γ between the individual effects of both players in one match is not standard. This covariance captures the idea that if $\mathcal{A}$ performs better on service than the rankings suggest, then one would expect that the probability that $\mathcal{B}$ will win a point on service is lower. It also captures, for example, fear against a specific opponent. Hence for our tennis data we expect (but do not impose) that η_a and η_b are negatively correlated ($\gamma < 0$).

The final assumption concerning quality is that the observed and unobserved parts are uncorrelated,

$$\operatorname{cov}(x_a, \eta_a) = \operatorname{cov}(x_a, \eta_b) = 0. \tag{4}$$

This is reasonable for our tennis application, because the rankings in x_a are determined well before the match starts (fixed at the end of the tournaments played in the week before Wimbledon). The assumption is also necessary, because otherwise the FGLS estimation procedure discussed in Section 3.4 will not be consistent (Kiviet 1995).

3.2 Dynamics

In contrast to the component Q_a, which captures the effects of variables that are known before the match begins, D_{at} captures the effects of variables that change during the match. We write

$$D_{at} = z_{at}'\delta \tag{5}$$

and interpret z_{at} as variables that reflect deviations from the iid assumption. In particular, previous points may influence the current point. This is *dependence*, and the variable $y_{a,t-1}$ is an obvious (be it simple) example. In addition, the current point may be played differently from other points. If this occurs, then we have *nonidentical distribution*, which could be measured by the importance of point t. In Section 4.2 we specify the dynamic regressors in such a way that deviations from the iid assumption (if present) are allowed to be heterogeneous across players.

In our application, the regressors z_{at} are completely determined by the history of the match up to point t and possibly exogenous attributes (such as the ranking of both players). We emphasize that the development of the match depends also on η_a and η_b, so that $\operatorname{cov}(z_{at}, \eta_a)$ and $\operatorname{cov}(z_{at}, \eta_b)$ may be nonzero.

3.3 Error Term: The Effect of Binary Structure

The third component in (1) is the error term ϵ_{at}. The binary structure of y_{at} has implications for the error term ϵ_{at}, particularly for its variance.

We assume that $E(\epsilon_{at}) = 0$. Regarding the second moments, we assume that

$$\begin{aligned}
&\operatorname{cov}(\epsilon_{at}, x_a) = \operatorname{cov}(\epsilon_{at}, x_b) = 0,\\
&\operatorname{cov}(\epsilon_{at}, \eta_a) = \operatorname{cov}(\epsilon_{at}, \eta_b) = 0,\\
&\operatorname{cov}(\epsilon_{at}, z_{as}) = 0 \quad (s \le t), \qquad \operatorname{cov}(\epsilon_{at}, z_{bs}) = 0 \quad (s \le S_{bt}),\\
&\operatorname{cov}(\epsilon_{at}, \epsilon_{as}) = 0 \quad (s \ne t), \qquad \operatorname{cov}(\epsilon_{at}, \epsilon_{bs}) = 0,
\end{aligned} \tag{6}$$

where S_{bt} denotes the total number of points served by $\mathcal{B}$ until the beginning of the current game (where $\mathcal{A}$ is serving). These are standard assumptions in the panel data literature. For our tennis dataset they are also reasonable, because the quality variables x_a and η_a are given at the beginning of the match and the dynamic regressors z_{at} depend only on the outcomes of the previous points.

In the panel data literature, it is usually assumed that the variance of ϵ_{at} is homoscedastic (Hsiao 1986, p. 33). In our case this is not possible, because of the binary character of the observations. Because $E(y_{at}) = E(y_{at}^2)$, we obtain

$$\begin{aligned}
\operatorname{var}(\epsilon_{at}) &= E\{(x_a'\beta + z_{at}'\delta)(1 - x_a'\beta - z_{at}'\delta)\}\\
&\quad - 2E\{(z_{at}'\delta)(\eta_a + \epsilon_{at})\} - \tau^2 \equiv \sigma_a^2,
\end{aligned} \tag{7}$$

so that $\operatorname{var}(\epsilon_{at})$ depends on a. Hence we must take heteroscedasticity into proper account.

3.4 Estimation

Assumptions (1)–(7) imply a dynamic binary panel model with random effects,

$$y_{at} = x_a'\beta + z_{at}'\delta + u_{at}, \qquad u_{at} = \eta_a + \epsilon_{at}, \tag{8}$$

and similarly for player $\mathcal{B}$. Stacking the $\{u_{at}\}$ into $T_a \times 1$ vectors u_a, and defining ι_a as the $T_a \times 1$ vector of 1's and I_{T_a} as

the $T_a \times T_a$ identity matrix, the $T \times T$ variance matrix of the error vector $(u'_a, u'_b)'$ of the whole match is given by

$$\Omega = \operatorname{var}\begin{pmatrix} u_a \\ u_b \end{pmatrix} = \begin{pmatrix} \sigma_a^2 I_{T_a} + \tau^2 \iota_a \iota'_a & \gamma \iota_a \iota'_b \\ \gamma \iota_b \iota'_a & \sigma_b^2 I_{T_b} + \tau^2 \iota_b \iota'_b \end{pmatrix}. \quad (9)$$

The matrix Ω is a generalization of the variance matrix of the standard two-error-components model and plays a crucial role in the FGLS estimation procedure discussed later. (Again, we simplify notation by writing Ω rather than Ω_{ab}, although Ω is different in each match.)

We wish to estimate the parameters of this model consistently. In dealing with consistency we always make the usual assumption that T_a (the number of points served by a player) is finite and that N (the number of matches) can become large. One could argue that T_a is large for our tennis data (see beginning of Sec. 3) and thus that the asymptotic approximation when $T_a \to \infty$ will work satisfactorily. But this is not the case, as we discuss in Section 5, question 8.

Estimation of a dynamic panel with random effects is no trivial exercise, mainly because z_{at} will be correlated with u_{at}, because variables such as $y_{a,t-1}$ in z_{at} will be correlated with η_a in u_{at}. Hence the regressors and the error term are contemporaneously correlated, implying that ordinary least squares (OLS) is inconsistent (even for large T_a).

The within-group or least-squares dummy variables (LSDV) estimator does take into account η, but is also inappropriate in this dynamic context. If we average over points for each player and subtract from (8), then we obtain $y_{at} - \bar{y}_a = (z_{at} - \bar{z}_a)'\delta + (\epsilon_{at} - \bar{\epsilon}_a)$. We have lost η_a, but again there is contemporaneous correlation and hence inconsistency (Kiviet 1995; Nickell 1981).

In the Appendix we describe how the model's parameters can be consistently estimated by FGLS. We use FGLS rather than generalized method of moments (GMM) (Ahn and Schmidt 1997; Arellano and Bover 1995), because there is much evidence that FGLS works well in finite samples (Balestra and Nerlove 1966; Sevestre and Trognon 1985), whereas Kiviet (1995) found mixed results for the finite-sample behavior of GMM in such models.

4. APPLICATION TO THE WIMBLEDON DATA

4.1 Specification of Quality Variables x_a

The "quality" variables x_a should reflect the observed quality of player $\mathcal{A}$ versus player $\mathcal{B}$. We base the definition of observed quality on the ranking of both $\mathcal{A}$ and $\mathcal{B}$ according to the lists published just before the Wimbledon tournament by the Association of Tennis Professionals (for the men) and the Women's Tennis Association (for the women). These two lists contain the official rankings based on performances over the last 52 weeks, including last year's Wimbledon. The ranking of player $\mathcal{A}$ is denoted by RANK_a. Note that RANK_a can be 500 even though only 128 players participate in the tournament.

Direct use of the rankings is unsatisfactory, because quality in tennis is a pyramid; the difference between the top two players (ranked 1 and 2) is generally greater than between two players ranked 101 and 102. The pyramidal structure is also evident in the seeding system and plays a role in acquiring points for the official ranking lists. The pyramid is based on "expected round:" 8 for the player who is expected to win the final (round 7), 3 for a player who is expected to lose in round 3, and so on. A problem with expected round is that it does not distinguish between, for example, players seeded 9–16, because all of them are expected to lose in round 4.

Thus motivated, we propose a smoother measure of expected round by transforming the ranking of each player into a variable R,

$$R_a = 8 - \log_2(\text{RANK}_a), \quad (10)$$

where $\log_2 x$ denotes $\log x$ with base 2. For example, if $\text{RANK} = 3$, then $R = 6.42$, and if $\text{RANK} = 4$, then $R = 6.00$. Note that players with $\text{RANK} > 128$ are not expected to play at Wimbledon at all; for these players, $R < 1$. Also, R can be negative, but this causes no problems. The average value of R over all players is 2.57 (men) and 2.74 (women). We show the robustness of our test results with respect to definition (10) in Section 5, question 2.

What then should be included—apart from an intercept—in x_a? Obviously, a player scores more points on service against a weaker opponent than against a stronger opponent. Hence relative quality (the gap between the two players) $R_a - R_b$ matters. But absolute quality (the overall quality of the match) $R_a + R_b$ also may matter, because we know that more points on service are scored in a match between two strong players than in a match between two weaker players (Magnus and Klaassen 1999a). We thus write

$$x'_a = (1, (R_a - R_b), (R_a + R_b)), \quad (11)$$

and let $\beta = (\beta_0, \beta_1, \beta_2)'$ denote the corresponding vector of three unknown parameters. Finally, we center both $R_a - R_b$ and $R_a + R_b$ by subtracting their sample means. This makes the interpretation of β_0 more transparent, giving the observed quality for an average match.

4.2 Specification of Dynamic Variables z_{at}

Because the focus of this article is on testing the iid hypothesis, the dynamic variables should contain "dependence" variables and "nonidentical distribution" variables. In essence we capture the dependence aspect by the previous point $y_{a,t-1}$ and the nonidentical distribution aspect by the importance of the point. There are three subtleties, however. The first subtlety concerns the influence of the previous point. The point preceding the first point of a game is the final point of the previous service game of the current server. Hence in measuring the influence of the previous point, there is a difference between the first point (where the previous point is "long" ago) and other points in a game. We thus define a dummy variable d_{at}, which takes the value 1 if the current point t is the first point in a game and 0 otherwise. We then define

$$\tilde{y}_{a,t-1} = \begin{cases} 0 & \text{if } d_{at} = 1 \\ y_{a,t-1} & \text{if } d_{at} = 0. \end{cases} \quad (12)$$

Using both $\tilde{y}_{a,t-1}$ and d_{at} allows us to study the effect of the previous point without having to delete the first point in every game.

The second subtlety is the measurement of the importance of a point. The importance of the tth point served by $\mathcal{A}$, denoted by imp_{at}, is defined as the probability that $\mathcal{A}$ will win the match if he or she wins the current point minus the probability that $\mathcal{A}$ will win the match if he or she loses the current point. This definition was first suggested by Morris (1977).

In the calculation of the importance imp_{at}, we assume that associated with each match there are two prematch (fixed) probabilities p_a and p_b. We define p_a as the estimated prematch probability that $\mathcal{A}$ wins a point on service. In particular, $p_a \equiv x_a'\tilde{\beta}$, where $\tilde{\beta}$ is the estimate of β under the assumption that $\delta = 0$. We treat p_a as a constant, not an estimate. The prematch probability p_b is defined similarly. Given the structure of a Wimbledon match (best-of-three sets for women, best-of-five sets for men; tie-break at 6–6, no tie-break in the deciding set), given the two fixed prematch probabilities, and assuming that the points are iid, our computer program calculates exactly the probability that $\mathcal{A}$ wins the match at each point of the match, and hence imp_{at} can be calculated at each point. The average of imp_{at} over all points and players is .028 for men and .034 for women. The distribution of imp_{at} is skewed to the right, indicating that important points are rare.

There is an apparent contradiction in the fact that, in calculating the importance of a point, we assume that points are iid ($\delta = 0$), although this is just what we intend to test. As a result, importance could be misrepresented. But the question here is not whether points are iid, but rather what is the *effect* of the iid assumption on the importance measure. The procedure is justified because p_a provides a very good first-order approximation to the probability that $\mathcal{A}$ wins a service point, as the deviation from iid is small (Sec. 4.3). Having defined the three explanatory variables, we now write the dynamic part as $D_{at} = \delta_1\tilde{y}_{a,t-1} + \delta_2 d_{at} + \delta_3 imp_{at}$.

However, a third subtlety must be taken into account. It may be the case that possible deviations from the iid assumption are player-dependent. For example, top players may be more equable (more iid) than lesser players. We account for this type of heterogeneity by letting δ_i depend on the rankings,

$$\delta_i = \delta_{i0} + \delta_{i1}(R_a - R_b) + \delta_{i2}(R_a + R_b) \quad (i = 1, 2, 3). \quad (13)$$

Because $(R_a - R_b)$ and $(R_a + R_b)$ are centered (Sec. 4.1), δ_{i0} gives the effect for an average match.

This then leads to the definition of the dynamic regressors,

$$z_{at}' = (\tilde{y}_{a,t-1}x_a',\ d_{at}x_a',\ imp_{at}x_a'), \quad (14)$$

and the associated parameters, $\delta' = (\delta_{10}, \delta_{11}, \delta_{12}, \delta_{20}, \delta_{21}, \delta_{22}, \delta_{30}, \delta_{31}, \delta_{32})$. If $\delta_{10} > 0$, then there is positive dependence ("winning mood"), so that winning the previous point increases the probability of winning the current point. This dependence is smaller for players who are better than their opponent (if $\delta_{11} < 0$) and in matches between two good players (if $\delta_{12} < 0$). If $\delta_{30} < 0$, then it is more difficult for the server to win an important point. The better player in a match can neutralize this effect somewhat if $\delta_{31} > 0$. If $\delta_{32} > 0$, then the effect is also smaller in matches between two good players.

Our main interest is the iid hypothesis, which now takes the form $\delta = 0$. Also of interest is the "homogeneity" hypothesis, which tests whether deviations from iid (if present) are homogeneous among players. This second hypothesis takes the form $\delta_{i1} = \delta_{i2} = 0 \quad (i = 1, 2, 3)$.

4.3 Estimation Results and Test of the iid Hypothesis

We estimate the 14 unknown parameters (three β's, nine δ's, τ^2, and γ) of our model by FGLS as explained in Section 3.4, based on the Wimbledon data described in Section 2. The results (not reported) show that four of the nine δ parameters (the same four for men and women) are not significantly different from 0 ("significant" is always at the 5% level), both individually and simultaneously. (Wald tests yield p values of 96% for men and 19% for women.) The nonsignificant parameters are δ_{11} (previous point × quality difference) and all three δ parameters associated with the first-point-in-game dummy d_{at}. We delete these four δ parameters and obtain a reduced model containing 10 unknown parameters. This model forms the basis for all subsequent discussion. The estimates of the reduced model are presented in Table 1.

The main conclusion derived from Table 1 is that the iid hypothesis ($\delta_{10} = \delta_{12} = \delta_{30} = \delta_{31} = \delta_{32} = 0$) is strongly rejected with p values of .03% for men and .01% for women. (In the unrestricted model the iid test is also rejected with p values of .50% for men and .03% for women). In the next section we show that this rejection is robust. The lack of independence and the lack of identical distribution are about equally important in this rejection. The independence hypothesis ($\delta_{10} = \delta_{12} = 0$) is rejected with a p value of 1.7% for men and .3% for women, whereas the identical distribution hypothesis ($\delta_{30} = \delta_{31} = \delta_{32} = 0$) is rejected with a p value of 1.7% for men and .5% for women.

Player homogeneity ($\delta_{12} = \delta_{31} = \delta_{32} = 0$) is also rejected with p values of .2% for men and .1% for women. (In the unrestricted model the homogeneity test is rejected with p values of 1.9% for men and .15% for women.) Hence deviations from iid are player-dependent.

Winning the previous point has a positive effect on winning the current point for both men and women, because $\delta_{10} > 0$. The stronger the players, the weaker

Table 1. Estimation Results for the Wimbledon Data

	Men's singles		*Women's singles*	
Constant (β_0)	.6456	(.0040)	.5596	(.0050)
Quality difference (β_1)	.0106	(.0014)	.0198	(.0017)
Quality sum (β_2)	.0035	(.0014)	.0040	(.0017)
Previous point $\tilde{y}_{a,t-1}$				
× constant (δ_{10})	.0085	(.0041)	.0123	(.0058)
× quality sum (δ_{12})	−.0028	(.0014)	−.0052	(.0019)
Importance imp_{at}				
× constant (δ_{30})	−.1304	(.0666)	−.1752	(.0779)
× quality difference (δ_{31})	.0394	(.0262)	.1405	(.0493)
× quality sum (δ_{32})	.0533	(.0231)	.0137	(.0241)
Random effects				
Variance (τ^2)	.0031		.0026	
Correlation (γ/τ^2)	−.5353		−.8317	
Wald tests				
iid	22.9901	[.0003]	25.2179	[.0001]
Homogeneity	14.6005	[.0022]	16.1926	[.0010]

NOTE: Standard errors are in parentheses; p values are in square brackets.

this effect: $\delta_{12} < 0$. At important points, the server has a disadvantage: $\delta_{30} < 0$. The effect is weaker for high-level matches than for low-level matches ($\delta_{32} > 0$, although not significantly in the women's singles) or when the server is better than the receiver ($\delta_{31} > 0$, although not significantly for the men).

Observed quality matters. As expected, the relative quality of the two players ("quality difference") is more important than the absolute quality ("quality sum"): $\beta_1 > \beta_2$. Comparing men's and women's singles, equality of the β parameters is obviously rejected (mainly because β_0 is very different for men and women). We also see that β_1 is larger for the women than for the men, which corresponds to the fact that the difference in strength between top players and lesser players is greater in the women's singles than in the men's singles.

Even though the iid hypothesis is rejected, one might have expected a stronger rejection, because the sample is so large. Although 7 of the 10 δ estimates are significant, their t ratios are between 2 and 3, and hence individually only marginally significant. Maybe iid is rejected, not because the null hypothesis is false, but because the standard errors are so small, because of the large sample. To examine this possibility, we need a test for iid that does not depend on the standard errors. We thus consider the *signs* of the δ estimates (Berkson 1938) and see that all five estimates have the same sign for men and women (see Table 1). Because the men and women samples are independent, this has probability 1/32 if points are iid (assuming for simplicity that the elements of $\hat{\delta}$ are mutually uncorrelated for both men and women; taking the correlation of the estimates into account yields a probability of 3.3%, very close to 1/32). Hence, without using the effect of the large sample on the standard errors, we reject iid as well.

We also compare the *magnitudes* of the δ estimates between men and women. Equality of the δ parameters for these two independent samples is *not* rejected (p value $= 33.1\%$). This shows that even with our large sample, not every hypothesis is rejected, and that there is considerable consistency of the results between men's singles and women's singles. Both findings support our conclusion that points in tennis are not iid.

Our results also suggest that—even though iid is rejected—the assumption of iid in specific applications (such as forecasting) could be relatively harmless. The results of the next section support this statement.

5. DIAGNOSTICS AND SENSITIVITY ANALYSIS

Clearly, our model (like all models) is imperfect. In this section we investigate whether the underlying assumptions are justified and how sensitive the main focus of the article (testing the iid hypothesis) is to possible imperfections of the model. The first activity is called diagnostic testing; the second sensitivity analysis. We ask and answer 11 questions, some (or all) of which may already have occurred to the reader.

1: Is the use of the linear probability model justified? We have already commented [in Section 3, after (1)] on the differences between logit, probit, and the linear probability model and argued that the linear probability model is appropriate in our case. The main reason why the linear probability model could fail is if estimated probabilities fell outside the (0,1) interval. This is never the case. We already know that the estimate of $E(y_{at}) = \Pr(y_{at} = 1)$ lies in the interval (.55, .75) for men and (.42, .73) for women. Moreover, at each point, $\hat{y}_{at}$ lies in the interval (.37, .91) for men and (.32, .85) for women. Further, the estimate of $E(y_{at}y_{as})$ lies in the interval (.31, .57) for men and (.18, .53) for women for all $0 < |t-s| \leq 10$. For $|t-s| > 10$, not much will change, because the boundaries are very stable for different $t-s$. Finally, the estimated $E(y_{at}y_{bs})$ also falls in the (0,1) range for all players.

2: Is our basic quality measure appropriate? Our basic quality measure $R_a = \delta - \log_2 \text{RANK}_a$ was introduced in (10). We defended this choice on theoretical grounds. If in contrast we assume a linear basic quality measure $R_a = -\text{RANK}_a$, then the iid hypothesis is also rejected with p values of .17% for men and .07% for women.

3: Is unobserved quality relevant? Put differently, are random effects actually present, which translates to the null hypothesis $\eta_a = \eta_b = 0$? We use a Hausman test based on our FGLS estimator and the restricted FGLS estimator of β and δ. (The restriction is $\eta_a = \eta_b = 0$, implying a diagonal Ω matrix.) Under the null hypothesis, the restricted FGLS estimator is consistent and efficient. Under the alternative, FGLS is consistent and restricted FGLS is inconsistent. The usual Hausman test, based on the difference between the two estimators and the difference between the two variance matrices, has asymptotically a $\chi^2(8)$ distribution and takes the values 81.8 for men and 45.5 for women with p values below .01%. The null hypothesis is thus very firmly rejected, unobserved quality is definitely relevant, and random effects should play a role in the estimation procedure.

4: Is any autocorrelation left in the residuals? We consider the FGLS residuals weighted by $\widehat{\Omega}^{-1/2}$ and denote these by $\hat{v}_t$. We regress $\hat{v}_t$ on $\hat{v}_{t-k}$ for $k = 1, 2, \ldots$, which gives estimates of the correlations r_k (with standard errors). Because each game contains at least four points, it is natural to consider only r_1, r_2, and r_3. These are all insignificant for both men and women. Moreover a joint Wald test of $r_1 = r_2 = r_3 = 0$ gives p values of 62.4% for men and 28.2% for women. Hence there is no evidence of residual autocorrelation.

*5: Are the non-*iid *effects measured appropriately?* We use $y_{a,t-1}$ to test independence, and importance imp_{at} to test whether points are identically distributed. We could replace $y_{a,t-1}$ by some other measure of dependence, such as "outperformance" of the server in the current game with respect to the current set, defined as relative frequency of points won in current game minus relative frequency of points won in previous service games of current set. [Albert (1993), Albright (1993), and Stern and Morris (1993) discussed similar measures in baseball.] Similarly, we could replace imp_{at} with a breakpoint dummy. The iid hypothesis is rejected in both cases.

6: Are more regressors needed? We have experimented inter alia by adding $y_{a,t-2}$ with an associated dummy, various measures of outperformance of the server in the current game with respect to the current set, various measures of outperformance of the server in the current set with respect to previous sets, similar measures for the receiver in his or her service games, a breakpoint dummy, and separation of the importance of point-in-match (imp_{at}) into its three components: importance of point-in-game, game-in-set, and set-in-match (Morris

1977). Although occasionally the extra parameters are significant, the iid hypothesis is rejected in each of the extended models, showing the robustness of our result.

7: Should the model allow for nonlinearities? We already allow for one form of nonlinearity by making δ dependent on the quality of the player and his or her opponent, see (13). But we have also experimented by adding $y_{a,t-1} \times imp_{at}$ and imp_{at}^2. The coefficients of these two nonlinear terms are not jointly significant (p values of 86.0% for men and 22.5% for women), and including the extra terms does not alter the outcome of the iid test.

8: Are the results sensitive to the method of estimation? Our method of estimation is FGLS (Sec. 3.4). But other estimation methods are available. The Anderson–Hsiao method (Anderson and Hsiao, 1981) does not reject the iid hypothesis (p values are 24.1% for men and 17.1% for women), although the resulting estimates do not contradict our FGLS estimates. Both findings are not surprising, because it is well known that the Anderson–Hsiao method, though consistent, is not efficient. Hence when effects are small, like in our tennis hypothesis, the Anderson–Hsiao method may not detect them. This motivates our desire for more efficient methods such as FGLS.

One could argue that T_a is large for our tennis data and thus that the asymptotic approximation when $T_a \to \infty$ will work satisfactorily (see Sec. 3.4). For $T_a \to \infty$, the within-group estimator (also known as the LSDV estimator) is consistent (Hsiao 1986, p. 74). But using this estimator leads to an estimate of δ_{10} that is far from significant. Hence, although T_a is "large," it is not large enough to justify an asymptotic approximation for $T_a \to \infty$. Nevertheless, iid is still clearly rejected.

9: What happens if we do not weight the data? We explained in Section 2 that the data are weighted to account for the underrepresentation of weaker players in the sample. If we do not weight the data, then the iid test still rejects with p values of .6% for men and .9% for women.

10: How is the in-sample fit? The usual R^2 is misleading as a diagnostic, because we are dealing with binary variables (Maddala 1983, p. 38). Instead, we use the root mean squared error (RMSE) and the mean absolute error (MAE) as our measures of fit, and compare our model to the "iid model" (where all δ's are 0). Our model (the "non-iid model") appears to perform only marginally better (lower MAE) than the iid model, as shown in Table 2. As a reference point, we also estimate the "constant model" (in which the β's also are 0, except the constant). Our model and the iid model have a much better fit than the constant model. This means that including the ranking variables improves the fit substantially, whereas including the non-iid variables does so only marginally.

11: How is the out-of-sample fit? We analyze the out-of-sample fit by estimating our model based on three years and then predicting the other year. We can take the first three years and predict the fourth year, but we can also, for example, estimate the last three years and predict the first year. In this way we obtain four predictions. To simplify the presentation, we average across years. The conclusions are the same as for the in-sample analysis (see Table 2). It also appears that the role of outliers is very limited—these would yield a good in-sample fit, but a bad out-of-sample fit. This is another proof of the robustness of our results. Hence our non-iid model survives out-of-sample analysis. In fact, all diagnostics and sensitivity analyses show that the rejection of iid based on the estimated model is robust in many directions against deviations from the underlying assumptions.

6. CONCLUDING REMARKS

In this article we have used 86,298 points (481 matches) at Wimbledon 1992–1995 to investigate whether points in professional tennis are iid. We reject this hypothesis, and the rejection is robust. Winning the previous point has a positive effect on winning the current point, and at important points the server has a disadvantage. We have also shown that the deviations from iid depend on the quality of the players; the stronger a player, the smaller the deviation from the iid hypothesis. These results are the same for men and women. This suggests that players should be trained to "play every point as it comes."

Even though we have shown that points are not iid, we have also shown that the divergence from iid is small. Hence in many practical applications concerning tennis—such as predicting the winner of the match while the match is in progress—the iid hypothesis will still provide an good approximation *if we correct for the quality of the players.* This makes our study a useful starting point for future work on tennis. Both the large dataset and the accurate FGLS method of estimation are essential in detecting the small deviation from iid.

In addition to the empirical findings on tennis, we have provided a theoretical contribution to the estimation of discrete dynamic panel data models. Our proposed structure allows for a binary dependent variable, dynamic regressors, and random individual effects but nevertheless is easy and fast to estimate by feasible GLS. Because estimation of binary dynamic panels is still an unsolved problem in the classical statistical literature, this approach may find application outside tennis as well.

Table 2. In-Sample and Out-of-Sample Fits

	In-sample				*Out-of-sample*			
	Men's singles		*Women's singles*		*Men's singles*		*Women's singles*	
Model	*RMSE*	*MAE*	*RMSE*	*MAE*	*RMSE*	*MAE*	*RMSE*	*MAE*
Non-iid	.4777	.4557	.4936	.4865	.4778	.4558	.4938	.4866
iid	.4777	.4559	.4936	.4869	.4778	.4559	.4937	.4869
Constant	.4787	.4577	.4965	.4928	.4787	.4577	.4965	.4928

APPENDIX: FEASIBLE GENERALIZED LEAST SQUARES ESTIMATION

FGLS estimation consists of two stages: consistent estimation of (the parameters in) Ω given in (9), and GLS on (8) using the estimated Ω.

Consistent Estimation of Ω

To estimate Ω consistently, we need consistent estimates of β, δ, τ^2, and γ [see (9) and (7)]. These are obtained in three steps. In step 1 we take first differences in (8) to remove the random effect η_a,

$$y_{at} - y_{a,t-1} = (z'_{at} - z'_{a,t-1})\delta + (\epsilon_{at} - \epsilon_{a,t-1}), \qquad (A.1)$$

and estimate δ from this equation. Because z_{at} is correlated with $\epsilon_{a,t-1}$, we need to estimate δ by instrumental variables. We follow Anderson and Hsiao (1981) by using $z_{a,t-1}$ as an instrument for $z_{at} - z_{a,t-1}$, because it is uncorrelated with the error term and correlated with the explanatory variable (Arellano 1989; Judson and Owen 1999; Kiviet 1995). We thus obtain a consistent estimate $\hat{\delta}$ of δ.

In step 2 we use $\hat{\delta}$ to rewrite (8) as

$$y^*_{at} = x'_a\beta + u^*_{at}, \qquad y^*_{at} = y_{at} - z'_{at}\hat{\delta},$$
$$u^*_{at} = u_{at} - z'_{at}(\hat{\delta} - \delta). \qquad (A.2)$$

Regressing y^*_{at} on x_a in (A.2) yields a consistent estimate $\hat{\beta}$ of β (Anderson and Hsiao 1982).

Step 3 is motivated by Hsiao (1986, p. 89). We use the consistent estimates $\hat{\delta}$ and $\hat{\beta}$ to estimate the covariance parameters τ^2 and γ. Because $u_{at} = \eta_a + \epsilon_{at}$ and letting $\bar{u}_a = (1/T_a)\sum u_{at}$, we find that

$$E(\bar{u}_a^2) = \tau^2\left(1 - \frac{1}{T_a}\right) + \frac{1}{T_a}(\tau^2 + \sigma_a^2), \quad \gamma = E(\bar{u}_a\bar{u}_b), \quad (A.3)$$

where σ_a^2 is restricted by (7). We replace the errors u_{at} by the residuals $\hat{u}_{at} = y_{at} - x'_a\hat{\beta} - z'_{at}\hat{\delta}$ and estimate $\tau^2 + \sigma_a^2$ by replacing β, δ, and $\eta_a + \epsilon_{at}$ $(= u_{at})$ in (7) by their estimates and the expectations by the appropriate sample means. Averaging (A.3) over players then yields consistent estimates $\hat{\tau}^2$ and $\hat{\gamma}$. Subtracting $\hat{\tau}^2$ from the estimated $\tau^2 + \sigma_a^2$ yields $\hat{\sigma}_a^2$ for all players. We thus obtain the consistent $T \times T$ variance matrix $\widehat{\Omega}$ from (9).

Consistency of Generalized Least Squares

We next show that GLS (with known Ω) is consistent. Because OLS is inconsistent, the consistency of GLS is not trivial. It rests on two key ingredients: the absence of initial conditions and the fact that observed quality is not correlated with the random effect; see (4).

Regarding the first ingredient, in economic datasets it is typical that for small t, the dynamic regressors z_{at} depend on endogenous variables from before the sample period, such as y_{a0}. If such initial values are correlated with η_a, then GLS becomes inconsistent (Hsiao 1986, p. 88). Our case is different, because the process that generates the data coincides with the start of the actual data collection, so that z_{at} cannot depend on variables such as y_{a0}—there are no points before the beginning of the match. Hence inconsistency because of the existence of initial conditions does not occur here (Blundell and Bond 1998; Kiviet 1995; Nerlove and Balestra 1992; Sevestre and Trognon 1985).

If both ingredients are present, then Anderson and Hsiao (1982) and Hsiao (1986, p. 88) showed (under normality) that GLS is consistent, because GLS then equals maximum likelihood. Because of the binary structure of y_{at}, we do not assume normality. But it is obvious in our case that GLS equals normality-based pseudo–maximum likelihood. Gourieroux, Monfort, and Trognon (1984) showed that the pseudo–maximum likelihood estimator is consistent. Hence GLS is consistent as well.

Now that we know that $\widehat{\Omega}$ is consistent and that GLS is consistent, we obtain new consistent estimates of β and δ by performing GLS on (8), using $\widehat{\Omega}$ instead of Ω. These FGLS estimates will be more efficient than the ones obtained from (A.1) and (A.2). The efficiency gain is important; see Section 5, question 8.

As a refinement of the procedure, we use the FGLS estimates of β and δ to form a new estimate of Ω and continue this process until convergence. The estimation results are *not* sensitive to this iterative procedure.

[Received March 1998. Revised November 2000.]

REFERENCES

Ahn, S. C., and Schmidt, P. (1997), "Efficient Estimation of Dynamic Panel Data Models: Alternative Assumptions and Simplified Estimation," *Journal of Econometrics*, 76, 309–321.

Albert, J. H. (1993), Comment on "A Statistical Analysis of Hitting Streaks in Baseball" by S. C. Albright, *Journal of the American Statistical Association*, 88, 1184–1188.

Albert, J. H., and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

Albright, S. C. (1993), "A Statistical Analysis of Hitting Streaks in Baseball," *Journal of the American Statistical Association*, 88, 1175–1196.

Anderson, T. W., and Hsiao, C. (1981), "Estimation of Dynamic Models With Error Components," *Journal of the American Statistical Association*, 76, 598–606.

——— (1982), "Formulation and Estimation of Dynamic Models Using Panel Data," *Journal of Econometrics*, 18, 47–82.

Arellano, M. (1989), "A Note on the Anderson–Hsiao Estimator for Panel Data," *Economics Letters*, 31, 337–341.

Arellano, M., and Bover, O. (1995), "Another Look at the Instrumental Variable Estimation of Error-Components Models," *Journal of Econometrics*, 68, 29–51.

Balestra, P., and Nerlove, M. (1966), "Pooling Cross Section and Time Series Data in the Estimation of a Dynamic Model: The Demand for Natural Gas," *Econometrica*, 34, 585–612.

Baltagi, B. H. (1995), *Econometric Analysis of Panel Data*, Chichester, U.K.: Wiley.

Berkson, J. (1938), "Some Difficulties of Interpretation Encountered in the Application of the Chi-Squared Test," *Journal of the American Statistical Association*, 33, 526–536.

Blundell, R., and Bond, S. (1998), "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models," *Journal of Econometrics*, 87, 115–143.

Brown, W. O., and Sauer, R. D. (1993), Comment on "Does the Basketball Market Believe in the 'Hot Hand'?" by C. F. Camerer, *American Economic Review*, 83, 1377–1386.

Camerer, C. F. (1989), "Does the Basketball Market Believe in the 'Hot Hand'?," *American Economic Review*, 79, 1257–1261.

Chib, S., and Greenberg, E. (1994), "Bayes Inference in Regression Models With ARMA(p, q) Errors," *Journal of Econometrics*, 64, 183–206.

Croucher, J. S. (1981), "An Analysis of the First 100 Years of Wimbledon Tennis Finals," *Teaching Statistics*, 3, 72–75.

——— (1995), "Replaying the 1994 Wimbledon Men's Singles Final," *The New Zealand Statistician*, 30, 2–8.

Gilovich, T., Vallone, R., and Tversky, A. (1985), "The Hot Hand in Basketball: On the Misperception of Random Sequences," *Cognitive Psychology*, 17, 295–314.

Gourieroux, C., Monfort, A., and Trognon, A. (1984), "Pseudo Maximum Likelihood Methods: Theory," *Econometrica*, 52, 681–700.

Honoré, B. E., and Kyriazidou, E. (2000), "Panel Data Discrete Choice Models With Lagged Dependent Variables," *Econometrica*, 68, 839–874.

Hsiao, C. (1986), *Analysis of Panel Data*, Cambridge, U.K.: Cambridge University Press.

Jackson, D., and Mosurski, K. (1997), "Heavy Defeats in Tennis: Psychological Momentum or Random Effect," *Chance*, 10, 27–34.

Johnson, V. E., and Albert, J. H. (1999), *Ordinal Data Modeling*, New York: Springer-Verlag.

Judson, R. A., and Owen, A. L. (1999), "Estimating Dynamic Panel Data Models: A Practical Guide for Macroeconomists," *Economics Letters*, 65, 9–15.

Kiviet, J. F. (1995), "On Bias, Inconsistency, and Efficiency of Various Estimators in Dynamic Panel Data Models," *Journal of Econometrics*, 68, 53–78.

Larkey, P., Smith, R., and Kadane, J. (1989), "It's Okay to Believe in the 'Hot Hand'," *Chance*, 2, 35–37.

Lindsey, G. R. (1961), "The Progress of the Score During a Baseball Game," *Journal of the American Statistical Association*, 56, 703–728.

Maddala, G. S. (1983), *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge, U.K.: Cambridge University Press.

Magnus, J. R., and Klaassen, F. J. G. M. (1999a), "On the Advantage of Serving First in a Tennis Set: Four Years at Wimbledon," *The Statistician (Journal of the Royal Statistical Society*, Ser. D), 48, 247–256.

——— (1999b), "The Effect of New Balls in Tennis: Four Years at Wimbledon," *The Statistician (Journal of the Royal Statistical Society*, Ser. D), 48, 239–246.

——— (1999c), "The Final Set in a Tennis Match: Four Years at Wimbledon," *Journal of Applied Statistics*, 26, 461–468.

Mátyás, L., and Sevestre, P. (1992), *The Econometrics of Panel Data*, Dordrecht: Kluwer Academic.

Morris, C. (1977), "The Most Important Points in Tennis," in *Optimal Strategies in Sport*, eds. S. P. Ladany and R. E. Machol, Amsterdam: North-Holland, pp. 131–140.

Nerlove, M., and Balestra, P. (1992), "Formulation and Estimation of Econometric Models for Panel Data," in *The Econometrics of Panel Data*, eds. L. Mátyás and P. Sevestre, Dordrecht: Kluwer Academic.

Nickell, S. (1981), "Biases in Dynamic Models With Fixed Effects," *Econometrica*, 49, 1417–1426.

Sevestre, P., and Trognon, A. (1985), "A Note on Autoregressive Error Components Models," *Journal of Econometrics*, 28, 231–245.

Simon, W. (1971), "Back-to-the-Wall Effect?," *Science*, 174, 774–775.

——— (1977), "Back-to-the-Wall Effect: 1976 Perspective," in *Optimal Strategies in Sport*, eds. S. P. Ladany and R. E. Machol, Amsterdam: North-Holland, pp. 46–47.

Siwoff, S., Hirdt, S., and Hirdt, P. (1987), *The 1987 Elias Baseball Analyst*, New York: Collier.

Stern, H. S. (1995), "Who's Hot and Who's Not: Runs of Success and Failure in Sports," in *1995 Proceedings of the Section on Statistics in Sports, American Statistical Association*, pp. 26–35.

Stern, H. S., and Morris, C. N. (1993), Comment on "A Statistical Analysis of Hitting Streaks in Baseball" by S. C. Albright, *Journal of the American Statistical Association*, 88, 1189–1194.